Relevance of interest points for eye position prediction on videos

Alain Simac-Lejeune, Sophie Marat, Denis Pellerin, Patrick Lambert, Michèle Rombaut, Nathalie Guyader

Relevance of interest points for eye position prediction on videos

Alain Simac-Lejeune[1,2], Sophie Marat[1], Denis Pellerin[1], Patrick Lambert[2], Michèle Rombaut[1], and Nathalie Guyader[1]

[1] Gipsa-lab, 961 rue de la Houille Blanche, BP 46, F-38402 Grenoble Cedex, France
[2] Listic (Université de Savoie), BP 80439 74944 Annecy-le-Vieux Cedex, France

Abstract. This papers tests the relevance of interest points to predict eye movements of subjects when viewing video sequences freely. Moreover the papers compares the eye positions of subjects with interest maps obtained using two classical interest point detectors: one spatial and one space-time. We fund that in function of the video sequence, and more especially in function of the motion inside the sequence, the spatial or the space-time interest point detector is more or less relevant to predict eye movements.

Key words: spatial and space-time interest points, eye position

1 Introduction

Images contain a very large amount of data, and image analysis often begins by the selection of some "interest points" which are supposed to carry more information than the rest of the pixels. The definition of these interest points is quite subjective, and generally depends on the aim of the analysis.

In the case of still images, many interest point detectors had been proposed. They are based on the detection of points having a significant local variation of image intensities in different directions (corners, line endings, isolated points of maximum - or minimum - local intensity, etc.). The most popular one is probably the Harris detector [1], with scale adaptive versions ([2],[3]). Different Gaussian based detectors have also been proposed - LoG (Laplacian of Gaussian), DoG (Difference of Gaussians), DoH (Determinant of the Hessian). It can be noted that DoG are used in the definition of the well known SIFT (Scale Invariant Feature Transform) approach [4][5]. Successful applications of interest points have been proposed in image indexing [6], stereo matching [7], object recognition [4], etc.

Only a few interest point detectors had been defined in the case of moving images. Laptev [8] proposed a space-time interest point detector which is a temporal extension of the Harris detector. He used this detector for the recognition of human actions (walking, running, drinking, etc.) in movies. In [9], Dollar proposed a space-time detector based on a 2D spatial gaussian filter jointly used with a quadrature pair of 1D temporal Gabor filters. However, this approach is

limited to the detection of periodic motions, such as a bird flapping its wings. In [10], Scovanner et al. proposed a temporal extension of SIFT descriptor.

Parallel to the research about interest points, other researches have proposed models to predict where people look at when freely viewing static or dynamic images. Given an image or video sequence, these bottom-up saliency models compute saliency maps, which topographically encodes for conspicuity (or "saliency") at every location in the visual input [11]. The saliency is computed in two steps: first, the visual signal is split into different basic saliency maps that emphasize basic visual features as intensity, color, orientation, and second, the basic saliency maps are fuzzed together to create the master saliency map. This map emphasizes which elements of a visual scene are likely to attract the attention of human observers, and by consequence their gaze. The model saliency map is then evaluated using different metrics. These metrics are used to compare the model saliency maps with human eye movements when looking at the corresponding scenes ([12],[13]).

In the same way the visual saliency maps are compared with subject eye movements, in this papers, we test whether the interest points are related to human eye movements. The goal of this papers is to measure the similarity between the interest maps obtained with two successful interest point detectors, one static and one dynamic, and the human eye position density maps. More precisely, we focus on the specificity of these two detectors (spatial/space-time). The papers is organized as follows. Section 2 presents the two interest point detectors which are chosen in this work. The eye movement experiment and the evaluation method are detailed in section 3. A relevance analysis is described in section 4. Conclusions are given in section 5.

2 Selection and description of interest point detectors

In the case of still images, several works have compared the performances of different interest point detectors. In [14], Schmid et al. introduced two evaluation criteria: the repeatability rate and the information content. The repeatability rate evaluates the detector stability for a scene under different viewing conditions (five different changes were tested: viewpoint changes, scale changes, image blur, JPEG compression and illumination changes). The information content measures the distinctiveness of features. Those two criteria directly measure the quality of the feature for tasks such as image matching, object recognition and 3D reconstruction. Using these two criteria the Harris detector appears to be the best interest point detector. For this reason, this detector will be chosen in the following, either in its spatial and space-time forms.

2.1 Spatial Interest Points: Harris detector

In an image, Spatial Interest Points (denoted SIP in the following) can be defined as points with significant gradients in more than one direction. In [1], Harris et

al. proposed to find such points using a second moment matrix H defined, for a pixel (x, y) having intensity $I(x, y)$, by:

$$H(x,y) = \begin{pmatrix} \frac{\partial^2 I}{\partial x^2} & \frac{\partial^2 I}{\partial x \partial y} \\ \frac{\partial^2 I}{\partial x \partial y} & \frac{\partial^2 I}{\partial y^2} \end{pmatrix} \quad (1)$$

In practice, the image I is first smoothed using a Gaussian kernel $g(x, y, \sigma)$ where σ controls the spatial scale at which corners are detected.

To obtain SIP, Harris et al. proposed to use a feature extraction function entitled "*salience function*", defined by:

$$R(x,y) = \det(H(x,y)) - k \times trace(H(x,y))^2 \quad (2)$$

The parameter k is empirically adjusted between 0.04 and 0.15 (0.04 is chosen in the following). SIP correspond to high values of the salience function extracted using a thresholding step (the salience function being normalized between 0 and 255, typical threshold value is 150).

2.2 Space-Time Interest Points: Laptev detector

Laptev et al. [15] proposed a spatio-temporal extension of the Harris detector to detect what they call "Space-Time Interest Points", denoted STIP in the following. STIP are points which are both relevant in space and time. These points are specially interesting because they focus information initially contained in thousands of pixels on a few specific points which can be related to spatio-temporal events in an image.
STIP detection is performed by using the Hessian-Laplace matrix H defined, for a pixel (x, y) at time t having intensity $I(x, y, t)$, by:

$$H(x,y,t) = \begin{pmatrix} \frac{\partial^2 I}{\partial x^2} & \frac{\partial^2 I}{\partial x \partial y} & \frac{\partial^2 I}{\partial x \partial t} \\ \frac{\partial^2 I}{\partial x \partial y} & \frac{\partial^2 I}{\partial y^2} & \frac{\partial^2 I}{\partial y \partial t} \\ \frac{\partial^2 I}{\partial x \partial t} & \frac{\partial^2 I}{\partial y \partial t} & \frac{\partial^2 I}{\partial t^2} \end{pmatrix} \quad (3)$$

As with the Harris detector, a gaussian smoothing is applied both in spatial and temporal domain. Two parameters σ_s and σ_t, one for each domain, control the spatial and temporal scale at which corners are detected. Typical values of σ_s and σ_t are respectively 1.5 and 1.2. In order to highlight STIP, different criteria have been proposed. As in [8], we have chose the spatio-temporal extension of the Harris corner function, entitled "*salience function*", defined by:

$$R(x,y,t) = \det(H(x,y,t)) - k \times trace(H(x,y,t))^3 \quad (4)$$

where k is a parameter empirically adjusted at 0.04 as for SIP detection. STIP also correspond to high values of the salience function R and are obtained using a thresholding step.

3 Eye position experiment and comparison metric

Eye positions are usually used to evaluate saliency models. Most of these models are inspired by the biology of the human visual system especially the processing of the retina and the primary visual cortex, ([11],[16],[17]). As the interest point models, these saliency models are based on low level properties of the stimuli. As the aim of this papers is to test whether the interest points are related to human eye movements, an experiment had been carried out in order to get the eye positions of subjects on particular video databases.

3.1 Experiment

We recorded the eye positions of fifteen subjects when they were viewing a video freely. This experiment was inspired by an experiment of Carmi and Itti [18] and is explained in more detail in [17]. Fifty three videos (25 fps, 720x576 pixels per frame) are selected from heterogeneous sources. The 53 videos are cut in small snippets of 1-3 seconds (1.86 ± 0.61), that are strung together to make up 20 clips of 30 seconds. Each clip contains at most one clip snippet from each continuous source. The total amount of snippets over the 20 clips is 305. The eye data are recorded using an Eyelink II eye tracker (SR Research), recording eye positions at 500Hz. As the frame rate is 25Hz, for each frame we compute the median of the 20 values related to this frame in order to get an eye position for each subject and for each frame. To relax the constraint on the exact positions, a 2-dimensional gaussian filtering is applied on each eye position point to obtain the human eye position density map M_h.

3.2 Comparison metric

The human eye position density map has to be compared to the interest maps provided by the interest point detectors. By looking what is done for saliency model evaluation, we can used several metrics ([12],[13]). In this papers we use the Normalized Scanpath Saliency (NSS) [19],[20] that was especially designed to compare eye positions and the salient locations emphasized by a saliency map M_m, and so it can be easily interpreted. The NSS is defined as follows:

$$NSS(t) = \frac{\overline{M_h(x,y,t) \times M_m(x,y,t)} - \overline{M_m(x,y,t)}}{\sigma_{M_m(x,y,t)}} \tag{5}$$

where $M_h(x,y,t)$ is the human eye position density map normalized to unit mean and $M_m(x,y,t)$ a saliency map for a frame t. $\overline{M_m(x,y,t)}$ and $\sigma_{M_m(x,y,t)}$ are respectively the average and the standard deviation of the model map $M_m(x,y,t)$.

In this papers the NSS is chosen to compare subjects' eye position and an interest map obtained with the interest-points detector as described below. The NSS is null if there is no link between eye positions and interest regions, negative if eye positions tend to be on non-interest regions and positive if eye positions

tend to be on interest regions. To sum up, a interest point model would be a good predictor of human eye fixations if the corresponding NSS value would be positive and high.

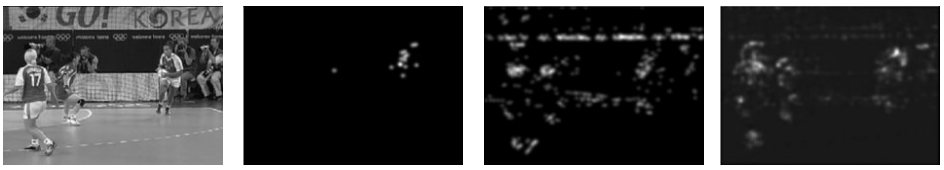
4 Relevance analysis

4.1 Eye position density map and interest maps

We chose to transform the sets of points (given by the eye position experiment and the interest point detectors) into maps by applying a 2D spatial gaussian filter on each point. This filtering allows to take the imprecision and the density of measures into account. For each frame of the different snippets presented in the previous section, three maps are worked out :

- Human eye position density map (M_h): which is obtained by applying a 2D Gaussian filtering on each eye position point. In this papers, it corresponds to the reference map.
- SIP interest map (M_{SIP}): this map corresponds to the SIP detector. As for the previous map, it is obtained by applying the same 2D Gaussian filtering on the SIP points.
- STIP interest map (M_{STIP}): this map corresponds to the STIP detector. For this map, we directly use a normalized version of the salience function $R(x, y, t)$ on which a 2D Gaussian filter is applied.

Figure 1 gives an example of these different maps. Human eye position density map ((M_h), fig. 1.b) does not look like the two different interest maps ((M_{SIP}),(M_{STIP}), fig. 1.c/d). There are very few highlighted regions on (M_h) compared to the interest maps. The more the highlighted regions of (M_h) will be also highlighted on the different interest maps, the more the NSS will be high.



a - Original image ("hand ball") | b - human eye position density map (M_h) | c - SIP interest map (M_{SIP}) | d - STIP interest map (M_{STIP})

Fig. 1. Example of maps extracted from a snippet

In order to determine the relevance of M_{SIP} and M_{STIP} maps according to the human eye position density map M_h, the NSS is calculated for each interest map. In the following, NSS_{SIP} (resp. NSS_{STIP}), denotes the NSS values obtained with M_{SIP} (resp. M_{STIP}).

4.2 Analysis on the global database

A temporal analysis of the NSS criteria is realized. Figure 2 shows the evolution of average NSS over time (or frames) of each snippet (see section 3.1).

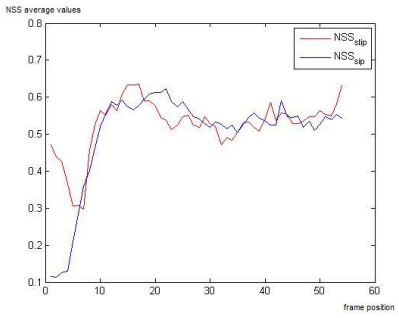


Fig. 2. NSS variations averaged over the 305 snippets over time plotted for the 55 first frames of each snippets. (NSS Average on the global database: NSS_{SIP} 0.50 - NSS_{STIP} 0.54)

First of all, both for SIP and STIP, we can note that the NSS values are positive which means that the interest points are relevant for eye position prediction. The second observation is related to the beginning of the evolution. The two curves present similar aspects after the first ten frames but are quite different at the beginning. That can be explained by the fact that after a shot cut between two snippets, humans' gaze stay at the same position, corresponding to the previous shot, for a short period before going to an interesting region of the new shot. As SIP interest map highlights interest points in a static way, after a shot cut the interest points change immediately and consequently are different from the regions gazed at. Thus the NSS is low. After a small delay, the subjects gazed at regions highlighted on the SIP map, and then the NSS increases. On the contrary, as the STIP interest map is built using a sliding window considering several frames before and after the current frame, during the first frames of a new snippet, STIP saliency map highlights interest points of the previous shot which are still gazed at by subjects.

It is particularly interesting to note that the NSS_{SIP} values are higher than the NSS_{STIP} values for approximately 65% of snippets. However, the $NSS_{STIP}average$ (0.54) is higher than $NSS_{SIP}average$ (0.50). Thus, when NSS_{STIP} is higher than NSS_{SIP}, it is significantly higher.

4.3 Analysis per semantical categories

We want to see if the interest points are more relevant for different semantical categories of snippets. An analysis shows that performances are different ac-

cording to the snippet content. Among the 305 snippets, we extract a number of classes of similar content. We have chosen to present four examples of interesting behavior: Traffic (18 snippets, 6% of the snippets), Team Sports (44 snippets, 14% of the snippets), Faces and/or Hands (47 snippets, 15% of the snippets) and Demonstration (30 snippets, 10% of the snippets). Figure 3 gives some images of snippets corresponding to these four classes.

a - Traffic b - Team sports c - Faces/hands d - Demonstration

Fig. 3. Image examples of the 4 classes

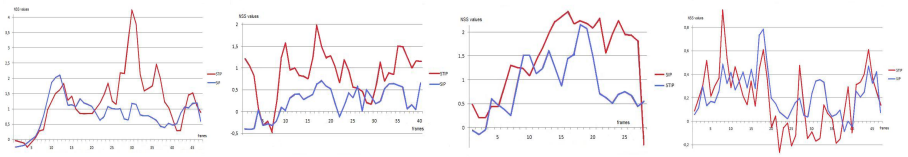
Table 1 summarizes the NSS values obtained for these different classes. This table shows the Traffic and Team Sports classes have got average and maximum values of NSS greater with STIP than with SIP. Furthermore, for Traffic class, the maximum value is very high. On the contrary, for the Faces/Hands class, SIP gives better average and maximum than STIP. Finally, NSS values for the demonstration class are close to 0, which means that there is a weak link between the eye positions and the interest points. To better understand these results, we give (Fig.4) the NSS evolution for an example of snippet of each class.

		Traffic	Team Sports	Faces/Hands	Demonstration
NSS_{SIP}	Average	0.86	0.17	1.85	0.19
	Maximum	2.10	0.72	4.78	0.78
NSS_{STIP}	Average	1.26	0.77	0.39	0.23
	Maximum	4.24	1.98	3.06	0.95
Number of snippets		18 (6%)	44 (14%)	47 (15%)	30 (10%)

Note that the minimum values of all the interest map is 0

Table 1. Average and maximum NSS values for different classes

Traffic class: This first example comes from the traffic class (fig 3.a), which presents sequences of traffic. Car traffic has special interest: it is characterized by a uniform movement, but with disorderly occasional variations more or less important: file changes, braking, accidents ... These discontinuities are particularly



a (Traffic) b (Team Sports) c (Faces/Hands) d (Demonstration)

Fig. 4. NSS over time for a snippet of each class

well detected and enhanced by STIP. In addition, these breaks easily attract visual attention. This explains quite well the correct values of the NSS and the improvement brought by temporal component compared to SIP (Figure 4.a).
On this figure, the NSS_{STIP} evolution exhibits a local maximum (value $\simeq 4$) near the thirtieth image. This peak corresponds to the abrupt change in direction of one of the vehicles.

Team sport class (using a ball): The second example concerns the category of team sports (fig3.b) using a ball: basketball, hand ball, rugby. These sports are characterized by rapid movements, rather erratic and with rapid changes. Furthermore, the more a player is close to the ball, or to the action, the more movements are rapid and disorderly. This context is favorable to STIP which tend to detect points with irregular motion.
Figure 4.b shows the evolution of NSS over images for SIP and STIP. Clearly, link with eye positions is better for STIP than for SIP. The local maxima of NSS_{STIP} generally correspond to sudden changes in the action. These changes attract eyes while providing a lot of STIP. So, for this class, the contribution of the temporal component to the interest point detection seems to be relevant. However, this result is not always true for football sequences. This counterexample is probably due to the fact that football images generally gives a wide view, which induces smoother motion.

Face/hand class: The third class is composed of close-up sequences of faces and hands (for instance a music concert - fig 3.c). This class represents the typical situations where the NSS_{SIP} is higher than the NSS_{STIP} (65% of global database).
In these sequences, the areas attracting the most attention are faces [21]. However, motion in these sequences is weak whereas they contain many spatial points of interest, generally located on faces or hands which are areas of visual attention. This explains the good results for SIP while adding the temporal component to interest point detection decreases the performance. Figure 4.c shows the evolution of NSS. NSS_{SIP} has a good level (average level approximately 1.5) and is almost all the time over NSS_{STIP}.

Demonstration/crowd class: The last class contains demonstrations or crowds (fig 3.d). The characteristics of this class is that movements are performed by a multitude of people covering almost all of the image.
Figure 4.d) shows that this class is characterized by a very low NSS relatively constant, around 0.2, as well for SIP as for STIP, even if in some cases STIP seems a little better than SIP.
The result indicates that the eye position do not match with the interest point. The problem is that in these sequences all people move which induces a lot of interest points uniformly distributed within images. But in the same way, visual attention is not captured by a particular area. Thus, the correspondence between eye positions and interest points is rare.

5 Conclusions and perspectives

The work presented in this papers has shown that interest points provided by specific detectors are globally relevant for eye position prediction. More precisely, we have studied the difference between spatial and space-time interest points, searching in which conditions interest points could be regarded as predictions of eye positions.

In order to analyse the relevance, the reference was built by recording eye position of subject which were compared to interest maps using the NSS metric. Experiment was run on a set of 305 snippets with various contents.

From the obtained results, we can get three main conclusions:

- globally, there is relevant link between eye positions and interest points (SIP and STIP). Hence interest points can be used as a prediction of gaze. The computational cost is very low regarding to other more dedicated methods.
- STIP provide a very good detection of eye positions when the sequence contains specials events, for examples: a car crash, somebody running and suddenly changing the direction of his run,
- On the contrary, when the semantic content is static (for faces and hands for example), the STIP do not work and SIP provide a very good detection of eye positions

A future extension of this work could be a collaborative use of SIP and STIP according to the video content. If information about the class or type of content is a priori known, the type of detector to use (SIP or STIP) can be easily chosen. If there is no additional information, intrinsic evaluation of STIP could help making this choice for optimum performance.

6 Acknowledgements

We thank the Rhône-Alpes region for its support with LIMA project.

References

1. C. Harris and M.J. Stephens, "A combined corner and edge detector," In Alvey Vision Conference, 1988.

2. T. Lindeberg, "Feature detection with automatic scale selection," International Journal of Computer Vision, pp. 77–116, 1998.
3. K. Mikolajczyk and C. Schmid, "Scale and affine invariant interest point detectors," International Journal of Computer Vision, pp. 63–86, 2004.
4. D.G. Lowe, "Object recognition from local scale-invariant features," International Conference on Computer Vision, pp. 1150–1157, 1999.
5. D.G. Lowe, "Distinctive image features from scale-invariant keypoints," International Journal of Computer Vision, pp. 91–110, 2004.
6. K. Mikolajczyk and C. Schmid, "Indexing based on scale invariant interest points," in Proc. ICCV, 2001, vol. 1, pp. 525–531.
7. T. Tuytelaars and L. Van Gool, "Wide baseline stereo matching based on local, affinely invariant regions," in British Machine Vision Conference, 2000, pp. 412–425.
8. I. Laptev, "On space-time interest points," International Journal of Computer Vision, vol. 64, no. 2/3, pp. 107–123, 2005.
9. P. Dollar, V. Rabaud, G. Cottrell, and S.J. Belongie, "Behavior recognition via sparse spatio-temporal features," in International Workshop on Performance Evaluation of Tracking and Surveillance, 2001, pp. 65–72.
10. S. Scovanner, P. Ali and M Shah, "A 3-dimensional sift descriptor and its application to action recognition," ACM Multimedia, 2007.
11. L. Itti, C. Koch, and E. Niebur, "A model of salincy-based visual attention for rapid scene analysis," IEEE Transaction on Pattern Analysis and Machine Intelligence, vol. 20, pp. 1254–1259, 1998.
12. B.W. Tatler, R.J. Baddeley, and I.D. Gilchrist, "Visual correlates of fixation selection : effects of scale and time," Vision Research, vol. 45, pp. 643–659, 2005.
13. A. Torralba, A. Oliva, M. S. Castelhano, and J.M. Henderson, "Contextual guidance of eye movements and attention in real-world scenes : The role of global features on object search," Psychological Review, vol. 113, no. 4, pp. 766–786, 2006.
14. C. Schmid, R. Mohr, and C. Bauckhage, "Evaluation of interest point detectors," International Journal of Computer Vision, vol. 37, no. 2, pp. 151–172, 2000.
15. I. Laptev and T. Lindeberg, "Space-time interest points," ICCV'03, pp. 432–439, 2003.
16. O. Le Meur, P. Le Callet, and D. Barba, "Predicting visual fixations on video based on low-level visual features," Vision Research, vol. 47, pp. 2483–2498, 2007.
17. S. Marat, T. Ho Phuoc, L. Granjon, N. Guyader, D. Pellerin, and A. Guérin-Dugué, "Modelling spatio-temporal saliency to predict gaze direction for short videos," International Journal of Computer Vision, vol. 82, no. 3, pp. 231–243, 2009.
18. R. Carmi and L. Itti, "Visual causes versus correlates of attentional selection in dynamic scenes," Vision Research, vol. 46, pp. 4333–4345, 2006.
19. R.J. Peters, A. Iyer, L. Itti, and C. Koch, "Components of bottom up gaze allocation in natural images," Vision Research, vol. 45, pp. 2397–2416, 2005.
20. R.J. Peters and L. Itti, "Applying computational tools to predict gaze direction in interactive visual environments," ACM Trans. On Applied Perception, vol. 5, no. 2, 2008.
21. M. Cerf, J. Harel, W. Einhauser, and C. Koch, "Predicting gaze using low-level saliency combined with face detection," Neural Information Processing System, 2007.